Conceptos Básicos de Colas

CONTENIDOS

1. Introducción
2. Elementos básicos de un modelo de colas
3. Notación Kendall: A / B / c / k / m / Z
4. Medidas de comportamiento. Sistema estable y sistema saturado
5. Ecuaciones de coste y fórmula(s) de Little
6. Comportamiento de transición y estacionario. Probabilidades límites. Propiedad PASTA

1. Introducción

•En numerosas situaciones de nuestra vida diaria esperamos en una cola o línea de espera,

como para comprar el billete del metro o la entrada de cine,
para cobrar un cheque en el banco,
para pagar en el supermercado o la cafetería,
para obtener una mesa en un restaurante,
para ser operado o atendido en un hospital,
para echar gasolina o pagar el peaje,
para desplazarnos en un atasco de tráfico, etc.

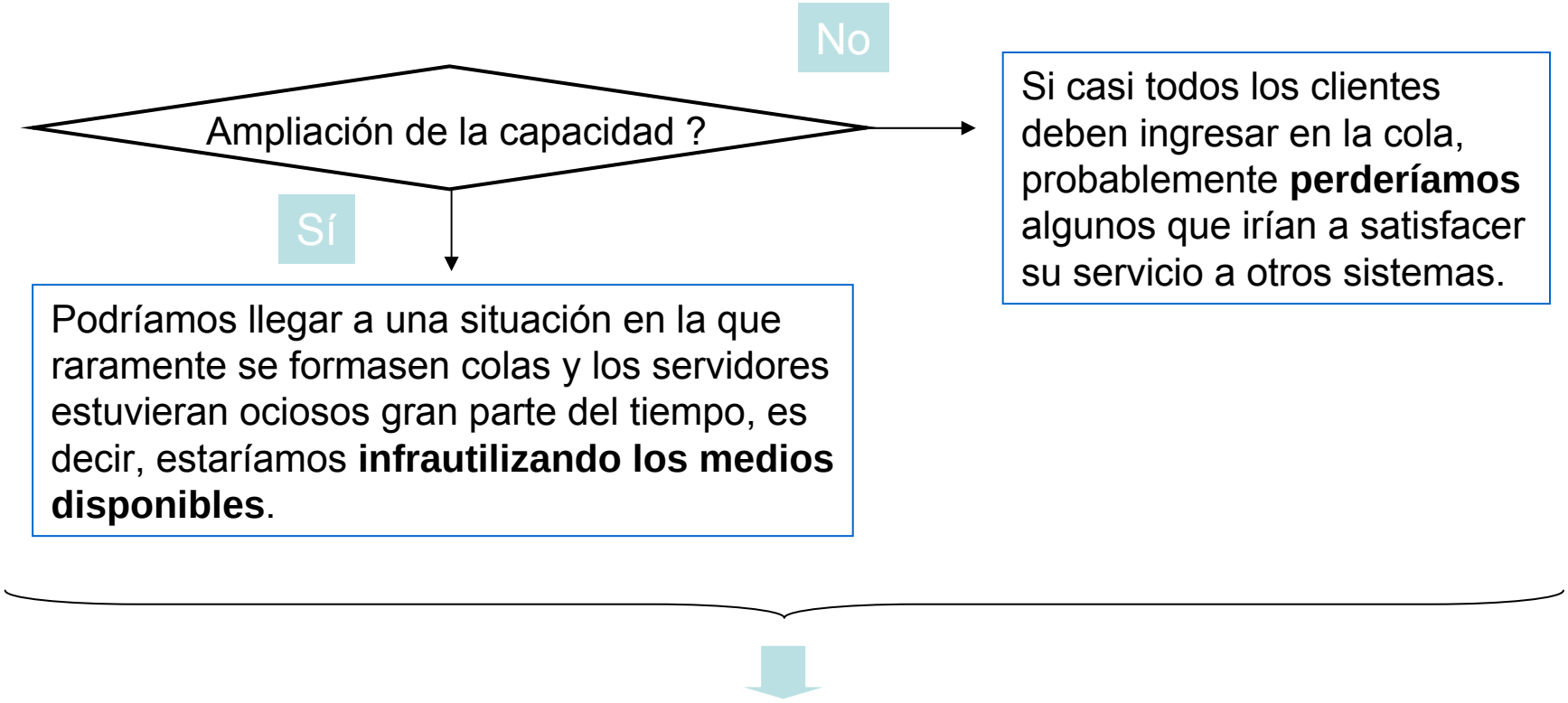
1. Introducción

•También en los sistemas informáticos son frecuentes los fenómenos de espera.

colas de personas esperando a usar un terminal,
colas de solicitudes de entrada/salida (E/S),
colas mensajes o paquetes de datos o programas informáticos que esperan para ser procesados por un sistema central o
colas llamadas telefónicas esperando una línea desocupada para completar la conexión.

1. Introducción

La espera se produce porque hay **más demanda de servicio que el disponible**. Sin embargo, ampliar esta capacidad de servicio no siempre es la solución adecuada.



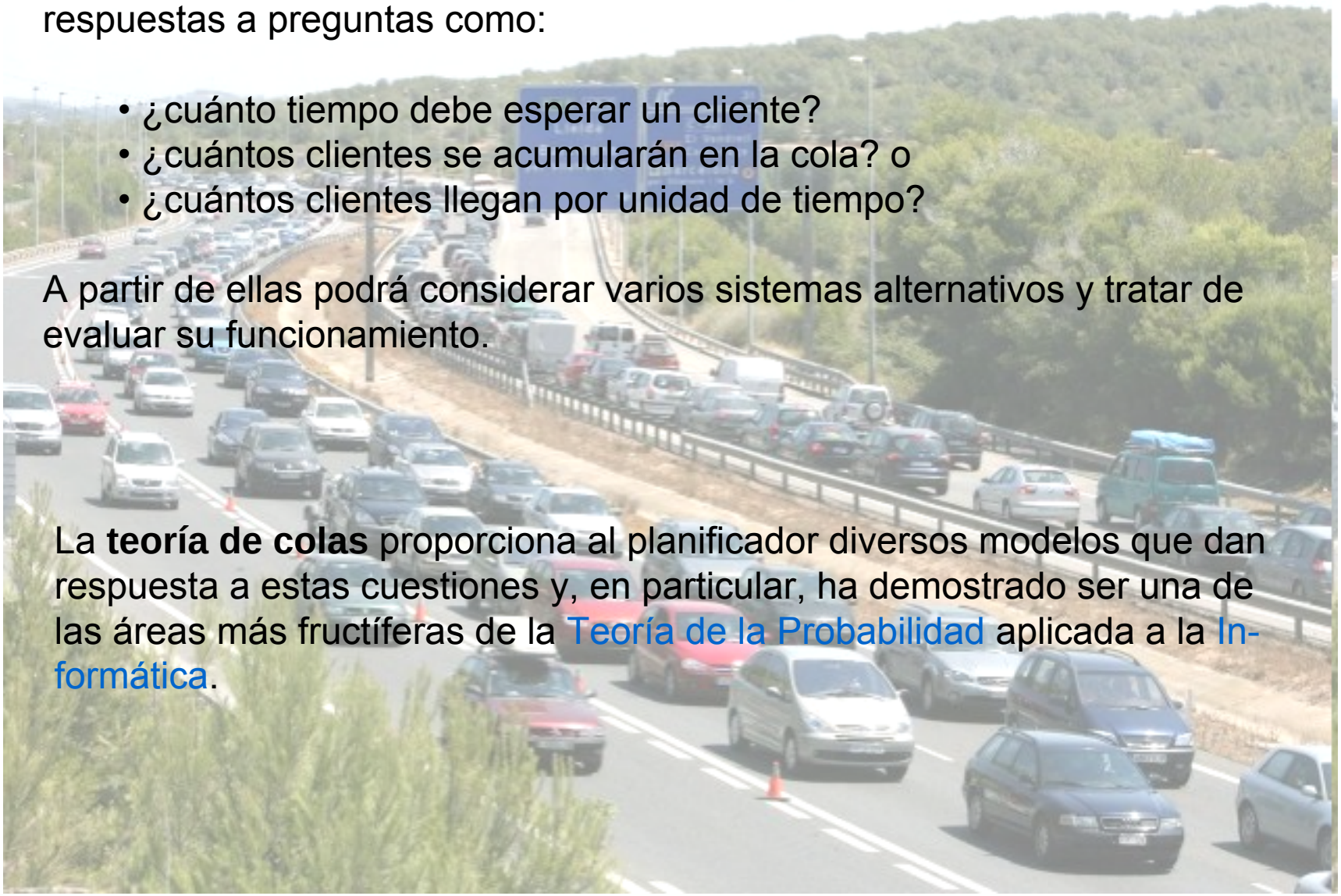
Por lo tanto, se trata de compensar un nivel adecuado de servicio con unos gastos no excesivos.

1. Introducción

Para llegar a una solución, el analista del sistema necesita conocer las respuestas a preguntas como:

- ¿cuánto tiempo debe esperar un cliente?
- ¿cuántos clientes se acumularán en la cola? o
- ¿cuántos clientes llegan por unidad de tiempo?

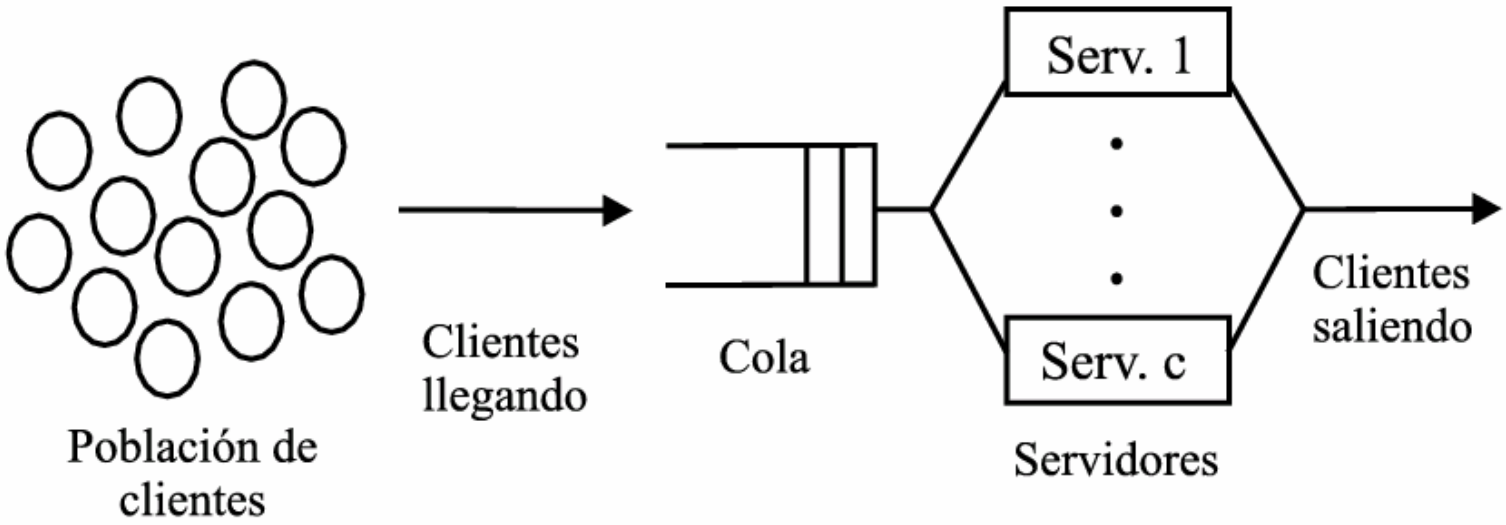
A partir de ellas podrá considerar varios sistemas alternativos y tratar de evaluar su funcionamiento.



La **teoría de colas** proporciona al planificador diversos modelos que dan respuesta a estas cuestiones y, en particular, ha demostrado ser una de las áreas más fructíferas de la Teoría de la Probabilidad aplicada a la Informática.

2. Elementos básicos de un modelo de colas

La siguiente figura esquematiza los elementos de un sistema de colas. Los clientes, que provienen de una población o fuente llegan al sistema para recibir algún tipo de servicio.



El dispositivo de servicio del sistema ofrece un conjunto (limitado) de servidores o recursos, a veces llamados canales, para satisfacer las peticiones de los clientes. Si cuando el cliente llega al sistema, todos los servidores están ocupados, deberá esperar en cola antes de empezar a recibir servicio. Una vez que el cliente recibe el servicio demandado abandona el sistema.

2. Elementos básicos de un modelo de colas

Una descripción más precisa de un sistema de colas requiere especificar en detalle 7 características básicas:

1. **Población o fuente de clientes**
2. **Modelo de llegadas**
3. **Modelo de servicio de cada servidor**
4. **Número de servidores o canales**
5. **Número de etapas de servicio**
6. **Capacidad del sistema**
7. **Disciplina de la cola**

2. Elementos básicos de un modelo de colas

1. Población o fuente de clientes

La población o fuente de clientes potenciales puede ser finita o infinita.

Infinita → conduce a sistemas con descripciones matemáticas más sencillas,
Finita → el número de clientes en el sistema afecta a la tasa de llegadas, que será cero si todos los clientes están en el sistema.

Si la población es finita pero suficientemente grande,
se asume que es infinita para simplificar el análisis del modelo.

2. Modelo de llegadas

Describe el patrón de llegadas de los clientes al sistema.

Si es **determinista** (llegadas están igualmente espaciadas en el tiempo) bastará caracterizarlo midiendo el número medio de llegadas por unidad de tiempo o el tiempo medio entre llegadas consecutivas.

Denotaremos con λ a la tasa media de llegadas o velocidad media, siendo por tanto 1/λ el tiempo medio entre llegadas.

2. Elementos básicos de un modelo de colas

En general, habrá **incertidumbre** en el modelo de llegadas y habrá que especificar la ley de probabilidad que rige el comportamiento aleatorio de las llegadas.

Suponemos que los tiempos de llegada de los clientes son

$$0 = t_0 < t_1 < t_2 < \ldots < t_n < \ldots$$

La observación del sistema comienza en el instante 0 y t_k es el instante en el que llega el cliente k-ésimo. Las variables aleatorias
$T_k = t_k - t_{k-1}$, k =1,2, 3,... representan los tiempos entre llegadas.

Normalmente, supondremos que el proceso estocástico continuo de tiempo discreto $T_1, T_2, \ldots$ es una secuencia de variables aleatorias independientes e idénticamente distribuidas (v.a.i.i.d.) con distribución T, con $E(T)=1/\lambda$.

El patrón de llegadas queda especificado dando la distribución de probabilidad de los tiempos de llegada o equivalentemente de los tiempos entre llegadas $P(T \leq t)$, que es la que suele utilizarse.

2. Elementos básicos de un modelo de colas

Los descriptores usuales son:

• *M*: tiempo entre llegadas es exponencial (el proceso de llegadas es de Poisson); la letra *M* proviene de la propiedad Markoviana de la exponencial;

• *D*: tiempo entre llegadas con patrón determinista o constante;

• E_k: tiempo entre llegadas con distribución de Erlang de *k* etapas;

• H_k: tiempo entre llegadas con distribución hiperexponencial de *k* etapas;

• *G*: tiempo entre llegadas sigue una distribución general o arbitraria.

Si el patrón de llegadas no cambia con el tiempo, es decir, la forma y los valores de los parámetros de la distribución son siempre iguales con el paso del tiempo, el modelo de llegadas es **estacionario**.

Posibilidad de que se produzcan llegadas en lotes o en masa al sistema, en lugar de un cliente cada vez. (Ej. llegada de familias a la consulta de un dentista). En este caso, el tamaño del lote es otra variable aleatoria del sistema.

2. Elementos básicos de un modelo de colas

Reacción del cliente al llegar al sistema:

- Un cliente puede decidir ingresar en la cola sin importarle la longitud de ésta, o si la considera demasiado larga puede decidir no entrar.

- También es posible que después de estar cierto tiempo en la cola decida marcharse.

- Por último, en el caso de disponer de dos o más colas simultáneamente, los clientes pueden decidir cambiarse de una a otra.

Estas situaciones son ejemplos de colas con clientes impacientes, y pueden considerarse llegadas dependientes del estado (o congestión) del sistema.

3. Modelo de servicio en cada servidor

Describe el tiempo de servicio que emplea un servidor en atender a un cliente.

2. Elementos básicos de un modelo de colas

En el caso **determinístico**, el patrón de servicio quedará descrito mediante el número de clientes servidos por canal por unidad de tiempo o mediante el tiempo requerido para servir a un cliente.

En caso contrario, será necesario especificar la distribución de probabilidad de la variable aleatoria s.
Supondremos que la secuencia de los tiempos de servicio de clientes sucesivos $s_1, s_2, \ldots$ también es un conjunto de v.a.i.i.d. con $E(s) = 1/\mu = W_s$.

Además, los procesos de llegadas y de servicio suelen considerarse independientes entre si.

Sin embargo, existe una diferencia importante entre ambos procesos. Cuando hablamos de tasa de servicio o tiempo de servicio, los términos están condicionados a que el sistema no esté vacío (el servidor esté ocupado).

La tasa de servicio es la capacidad intrínseca del servidor para satisfacer las necesidades de los clientes que, en general, es distinta de la tasa de salidas o número de clientes que dejan el canal de servicio una vez satisfechas sus necesidades.

2. Elementos básicos de un modelo de colas

Es decir, la tasa media de servicio es la tasa media a la que el servidor procesa a los clientes si estuviese ocupado el 100% del tiempo.

La notación para los patrones de servicio: M, D, E_k, H_k, G.

Así, M significa que la variable aleatoria s es exponencial de parámetro μ, que es la más usada. Por la propiedad de pérdida de memoria de la exponencial, el tiempo restante hasta completar el servicio de un cliente es independiente del tiempo que este cliente lleve en el canal.

El servicio puede ser **estacionario** o no respecto al tiempo (Ej. Un servidor que va aprendiendo y se vuelve más eficiente según adquiere experiencia).

La dependencia del tiempo no debe confundirse con **dependencia del estado**, es decir, del número de clientes que hay en el sistema. (Ej. trabajar más rápido según ve que se incrementa el número de clientes en la cola).

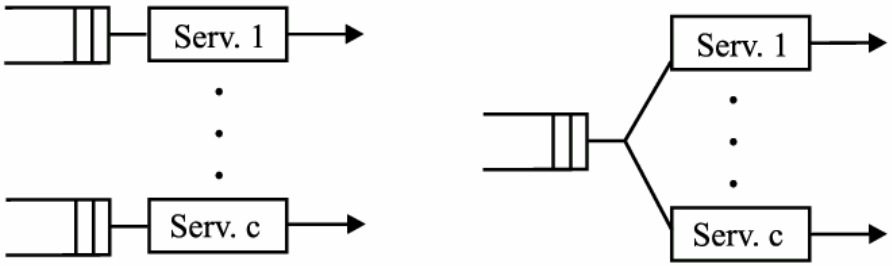
Puede haber situaciones en las que varios clientes sean atendidos simultáneamente por el mismo servidor, es decir, servicio por lotes o en masa, como un ordenador con procesamiento paralelo o turistas en una visita guiada.

2. Elementos básicos de un modelo de colas

4. Número de servidores o canales

El sistema de colas más sencillo tiene un único servidor, que atiende a un solo cliente cada vez. Un sistema multicanal o multiservicio dispone de *c* canales paralelos y puede dar servicio a *c* clientes a la vez.

La siguiente figura muestra dos variaciones de sistemas multicanal, cada canal tiene su propia línea de espera (cajas de supermercados o pasar la ITV de un coche), y una sola cola para todos los canales (turno en la peluquería).



Normalmente, se supone que los servidores son idénticos y funcionan de forma independiente unos de otros.
En el caso extremo de infinitos servidores, utilizado a veces como aproximación, cada cliente que llega es atendido inmediatamente.

2. Elementos básicos de un modelo de colas

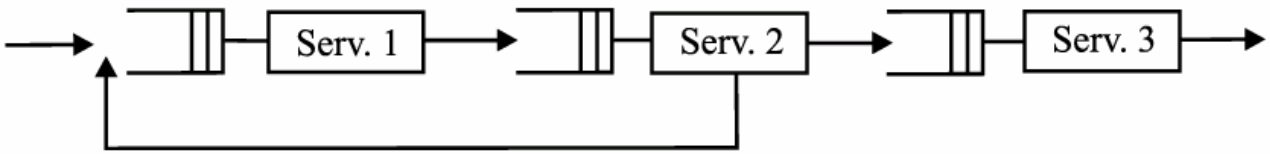
5. Número de etapas de servicio

A veces existen varias etapas de servicio por las que debe pasar el cliente, como en una cadena de producción o de montaje. Variantes:

- cada etapa acepta un cliente una vez que ha terminado el servicio del anterior (línea de sistemas de espera),

- la primera etapa sólo acepta un nuevo cliente cuando el anterior ha abandonado la última etapa (servidor constituido por diversas etapas).

La siguiente figura ilustra un sistema de colas con la primera variante, y además, presenta reciclado o retroalimentación, típico en procesos de fabricación con inspecciones para control de calidad en ciertas etapas. Un artículo que no cumple ciertas normas de calidad debe ser reprocesado.



2. Elementos básicos de un modelo de colas

6. Capacidad del sistema

En algunos sistemas de colas hay limitación física sobre el número máximo *K* de clientes que puede haber en el sistema. Cuando la cola alcanza cierta longitud, cualquier cliente que llega es rechazado hasta que se dispone de sitio, por completarse algún servicio.

El estado del sistema influye en la tasa de entradas al sistema λ_e, que es igual a la tasa de llegadas al sistema λ menos el número medio de clientes que no entran al mismo.

Un caso extremo son los llamados sistemas de pérdidas, que no admiten colas, como algunos sistemas de comunicación telefónica. En otros sistemas, sin embargo, puede suponerse que su capacidad es infinita y todo cliente que llega puede esperar hasta que se le proporcione servicio.

7. Disciplina de la cola

La disciplina de gestión de la cola o estrategia de servicio es la forma en que se seleccionan los clientes que aguardan en la cola para entrar en el dispositivo de servicio.

2. Elementos básicos de un modelo de colas

- FIFO (first-in, first-out) o FCFS (first-come, first-served). La más corriente es la de "el primero en llegar es el primero en entrar“. Es la que se supondrá por defecto.

- LIFO (last-in-first-out) o LCFS . El primero en entrar es el último en llegar. Usada en muchos sistemas de inventario en los que las unidades no son perecederas y resulta más sencillo tomar las unidades más cercanas, que se almacenaron más tarde.

- SIRO (service in random order). Servicio en orden aleatorio, independientemente del instante de llegada a la cola.

- SIFO (shortest-in, first-out) o SJF (shortest job first). Se sirve primero al cliente que demanda un tiempo de servicio menor.

- RR (round robin, turno robado). Reparte el tiempo del servidor equitativamente entre todos los clientes que esperan. Si el cliente no termina su servicio al final de la rodaja de tiempo que le corresponde utilizar, retorna a la cola, que se gestiona mediante disciplina FIFO. Esto se repite hasta que el cliente termina su servicio.

2. Elementos básicos de un modelo de colas

• PS (processor sharing) o de compartición del procesador. Disciplina RR en la que las rodajas de tiempo son infinitamente pequeñas. Es como si todos los clientes fueran servidos simultáneamente y sus tiempos de servicio incrementados de la misma forma.

Esquemas de prioridades → dan trato preferencial a ciertos clientes sobre otros. Los clientes están divididos en clases de prioridades. Los de prioridades más altas son servidos antes que los de prioridades más bajas, independientemente de su instante de llegada al sistema. Las prioridades pueden ir variando con el paso del tiempo.

- prioridad expulsiva, o con desalojo o apropiación (preemptive). Se interrumpe el servicio, el cliente recién llegado comienza a ser servido, y el cliente cuyo servicio ha sido interrumpido vuelve a la cabeza de la cola de su clase. Cuando el cliente desalojado reanude su servicio, éste comenzará desde el principio o desde el punto de interrupción, dependiendo del tipo de sistema.

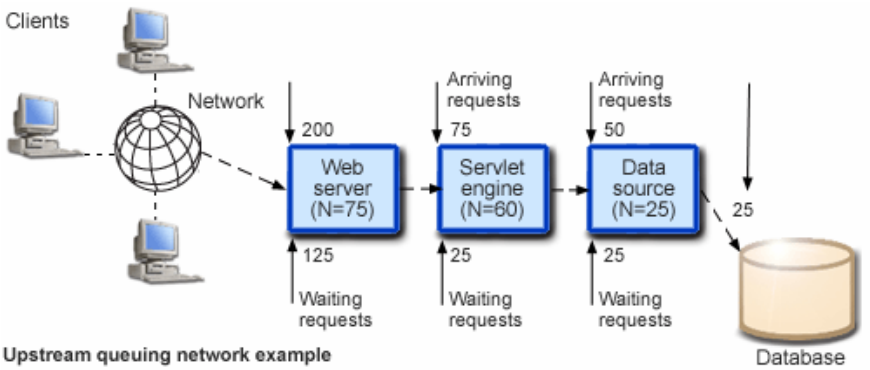
- prioridad sin desalojo, el cliente recién llegado espera hasta que el cliente siendo servido complete su servicio.

3. Notación Kendall: A / B / c / k / m / Z

Todos estos elementos básicos que describen un sistema de colas se representan mediante una notación abreviada estándar, denominada notación de Kendall (en honor a David Kendall).

Escribiremos *A/B/c/K/m/Z*, donde:

- *A* indica la distribución del tiempo entre llegadas,
- *B* la distribución del tiempo de servicio,
- *c* el número de canales de servicio (c≥1),
- *K* la capacidad del sistema,
- *m* el tamaño de la población y
- *Z* la disciplina de la cola.



Upstream queuing network example

Por ejemplo, *D*/*M*/3/40/∞/LIFO.

En muchos casos no hay límite sobre la capacidad del sistema, la fuente de clientes es infinita y la disciplina es FIFO. En estas situaciones suelen omitirse los tres símbolos finales. Así, M/G/4 es lo mismo que M/G/4/∞/∞/FIFO.